Supplementary Information for

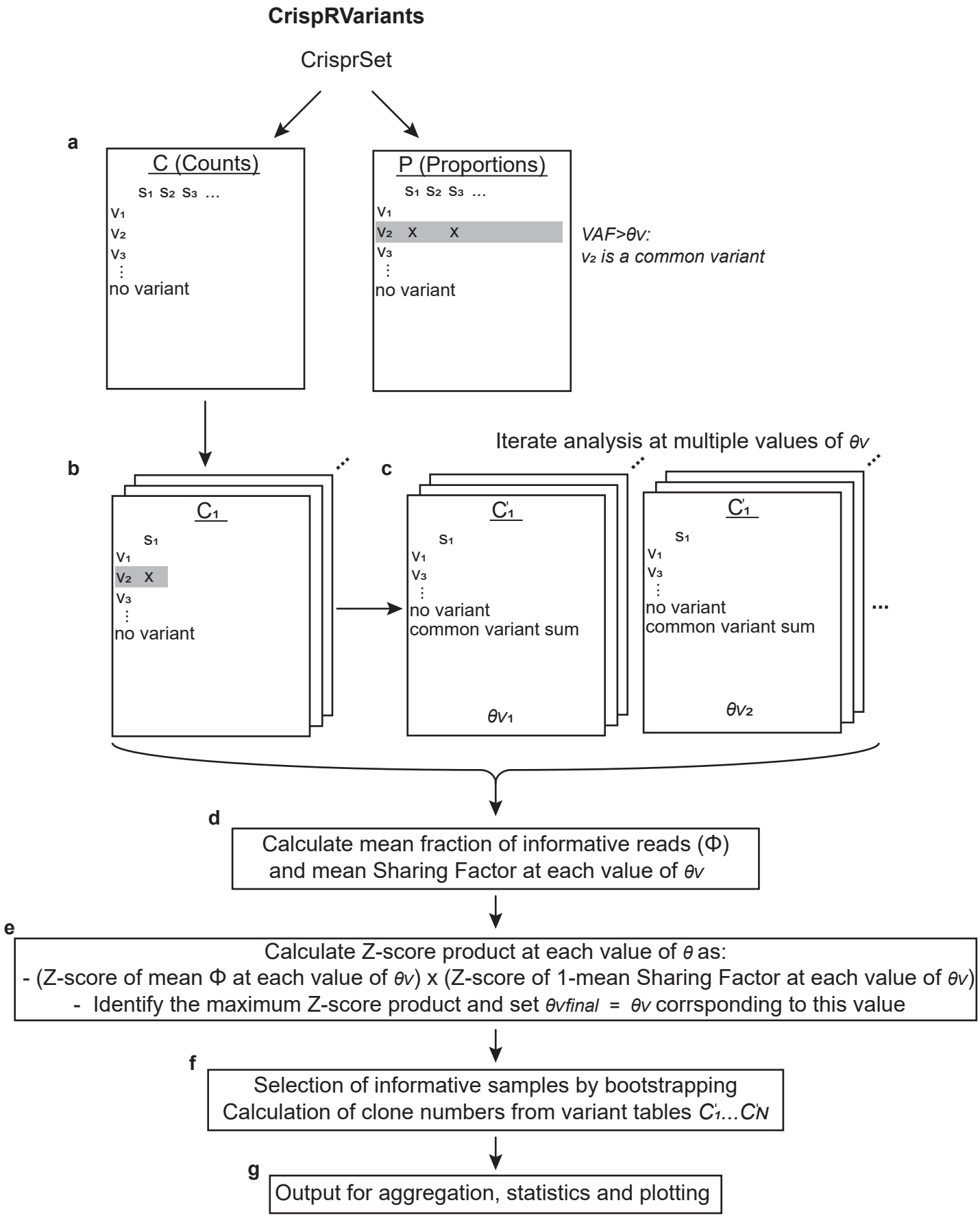
Quantifying Hematopoietic Stem Cell Clonal Diversity by Selecting Informative Amplicon Barcodes

Emily M. Teets[1†], Charles Gregory[1†], Jami Shaffer[1], James S. Blachly[1,2‡], Bradley W. Blaser[1*‡]

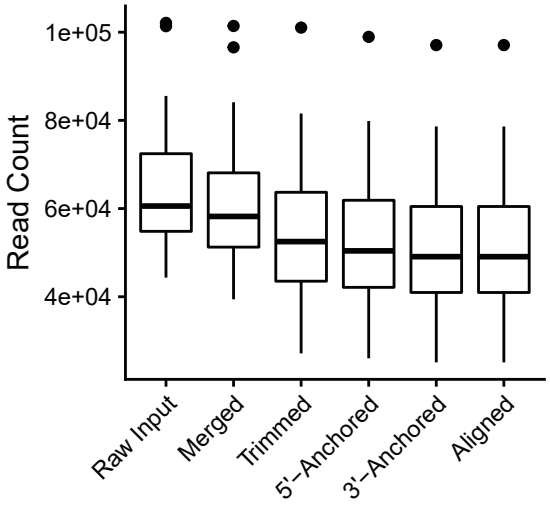
[1]The Ohio State University College of Medicine, Department of Medicine, Division of Hematology; The Ohio State University Comprehensive Cancer Center

[2]The Ohio State University College of Medicine, Department of Biomedical Informatics

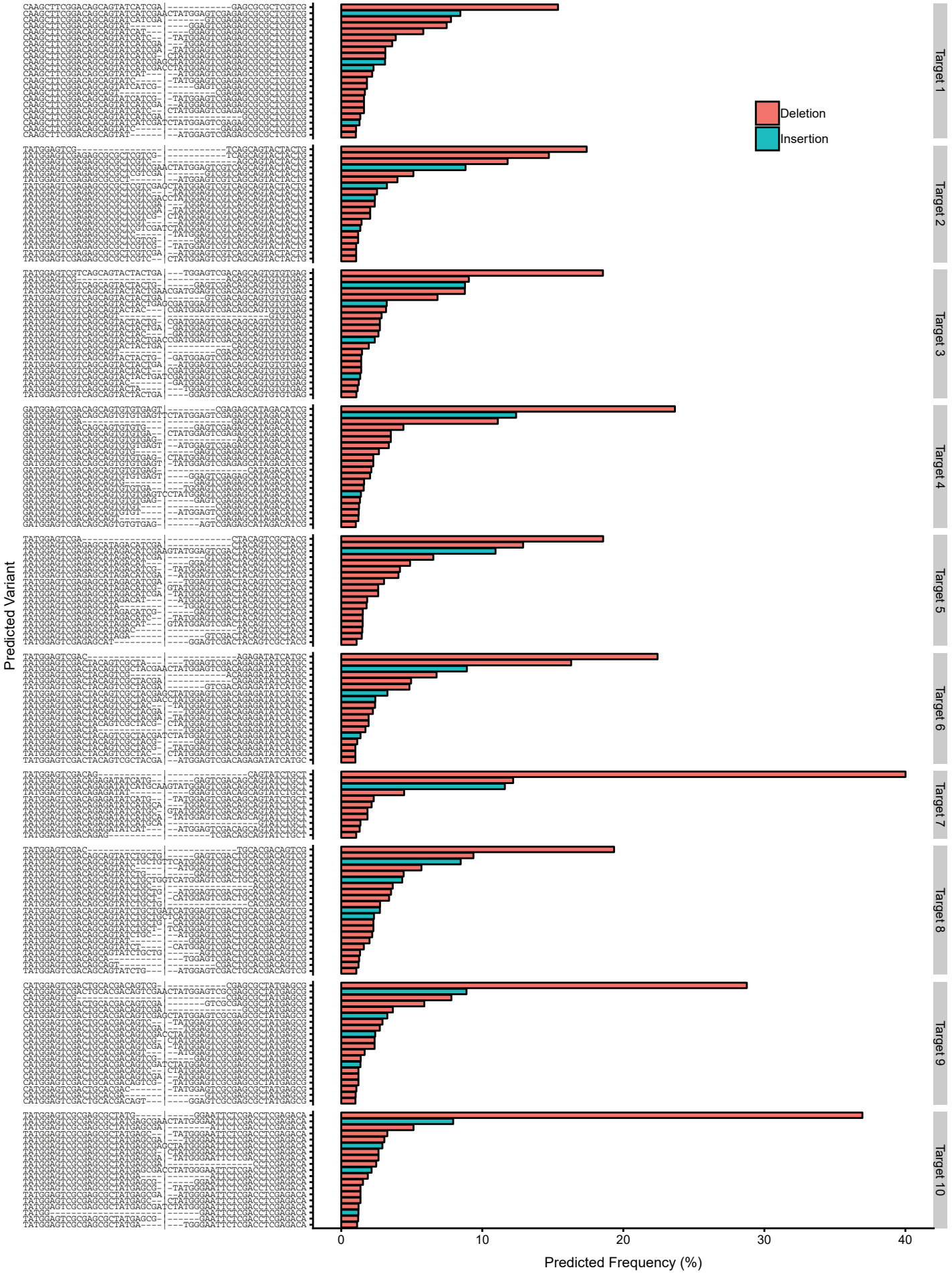
Supplementary Figure S1



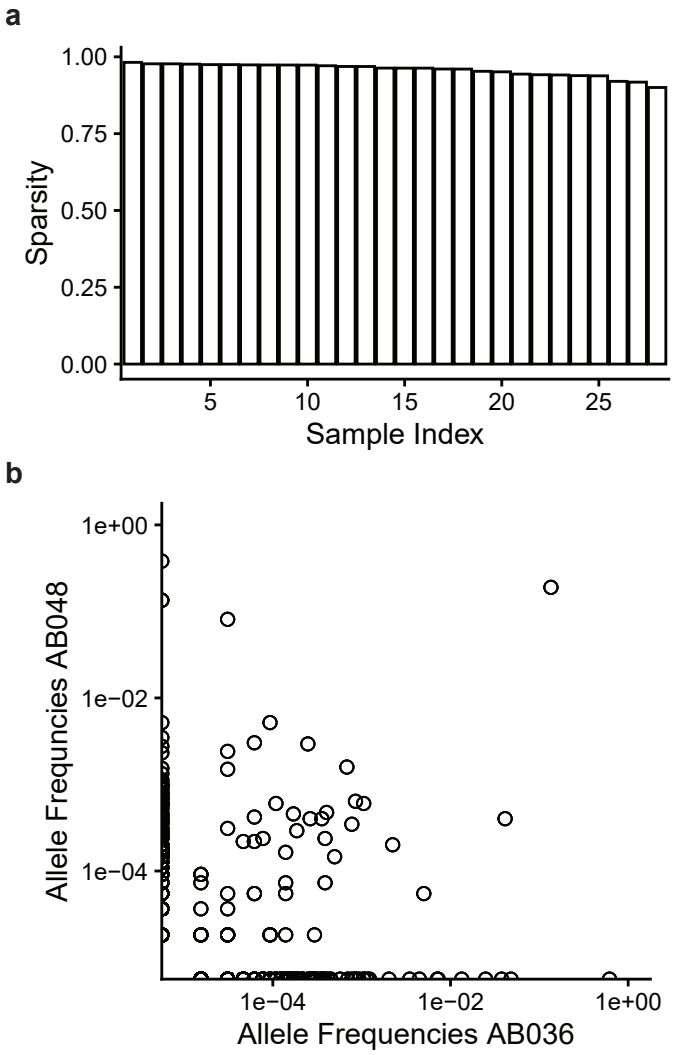
Supplementary Figure S2



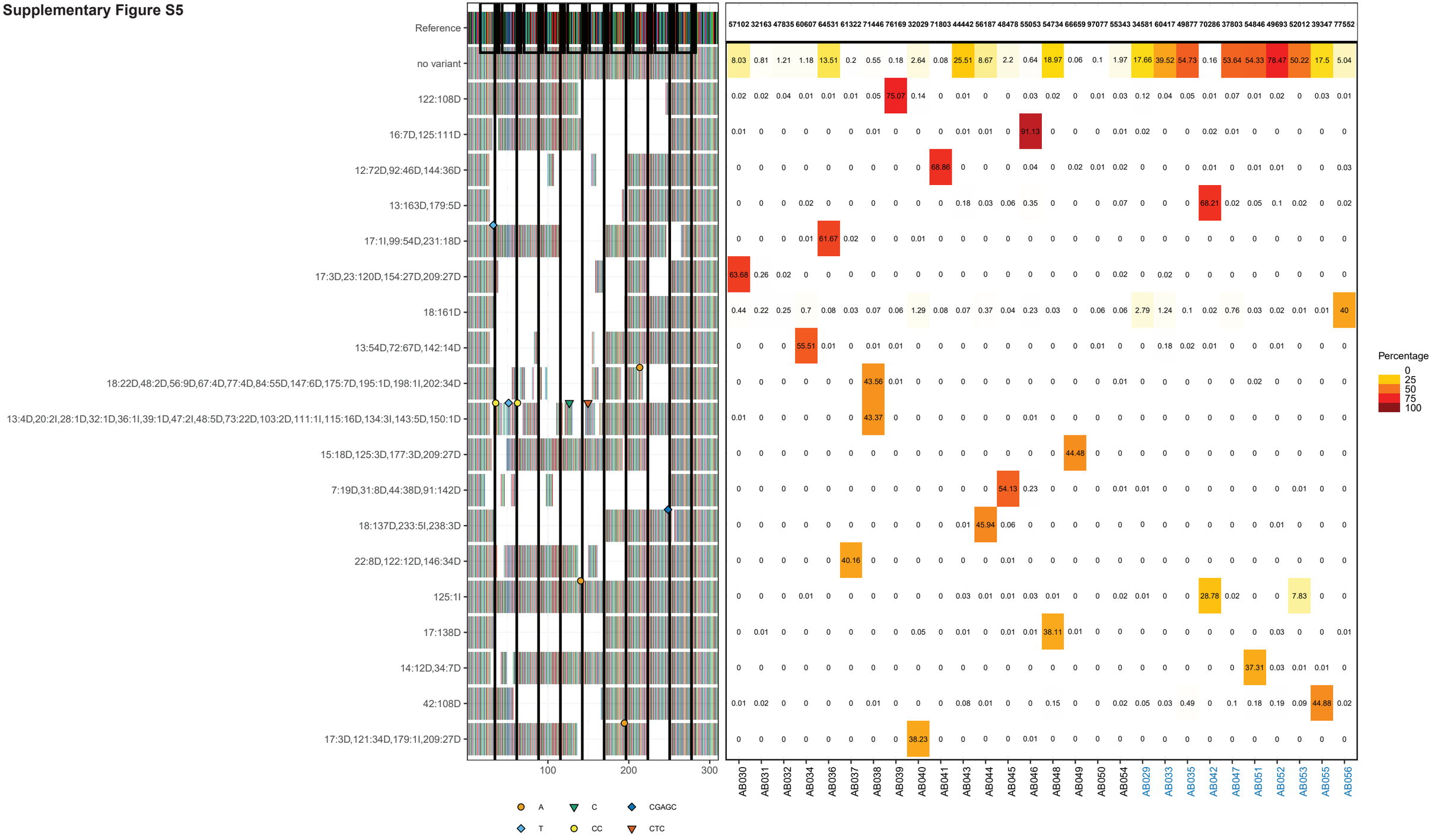
Supplementary Figure S3



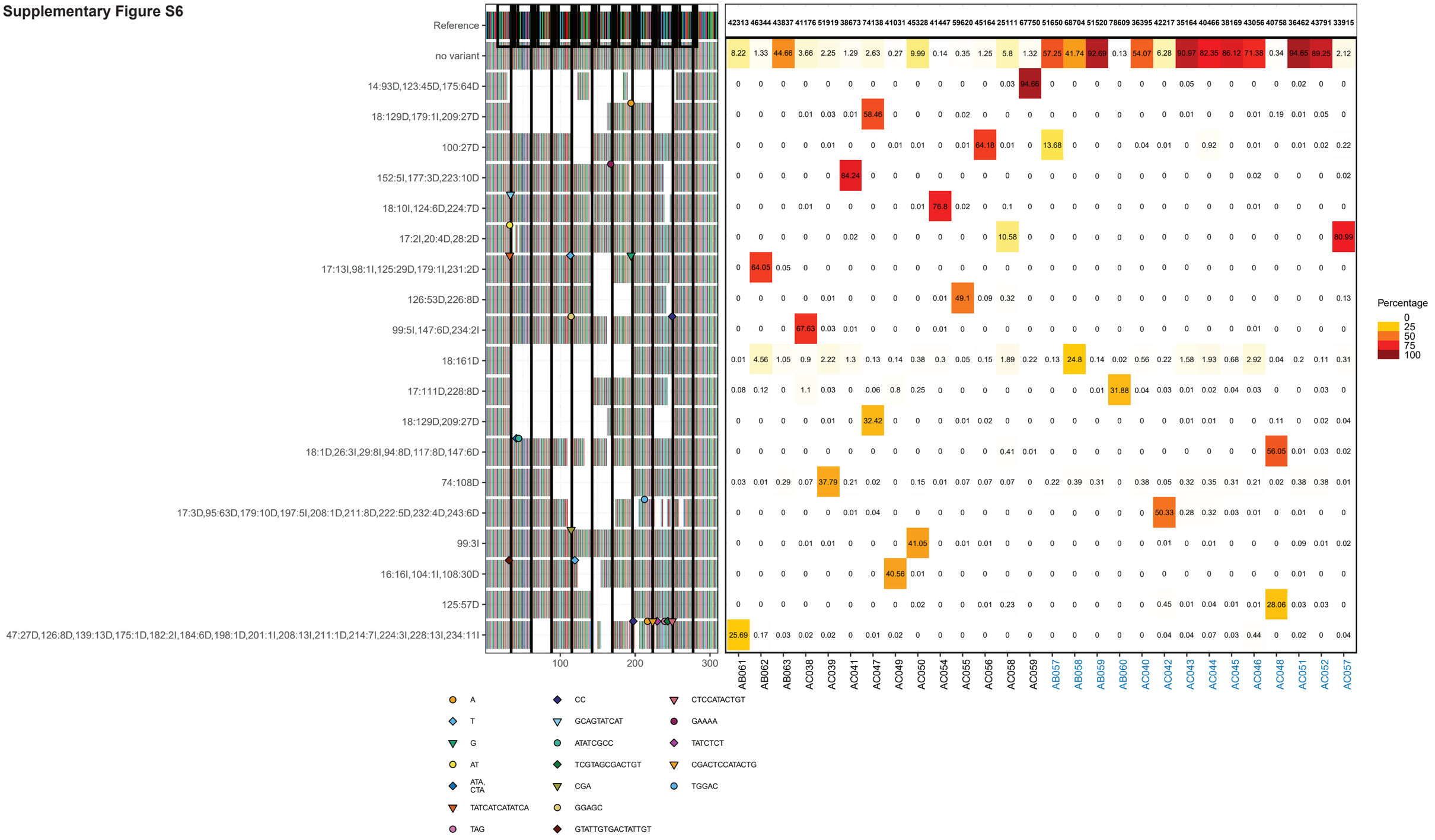
Supplementary Figure S4



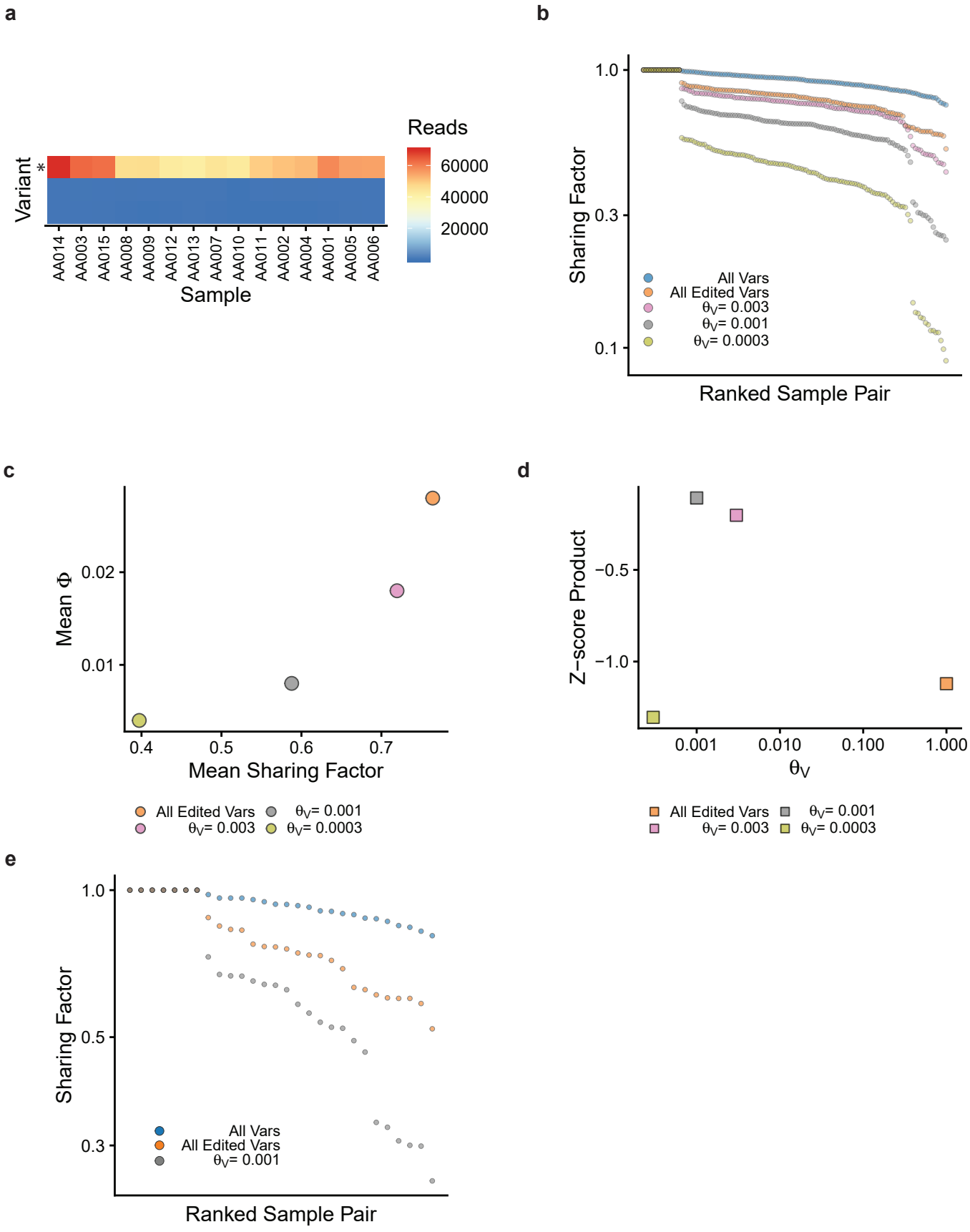
Supplementary Figure S5



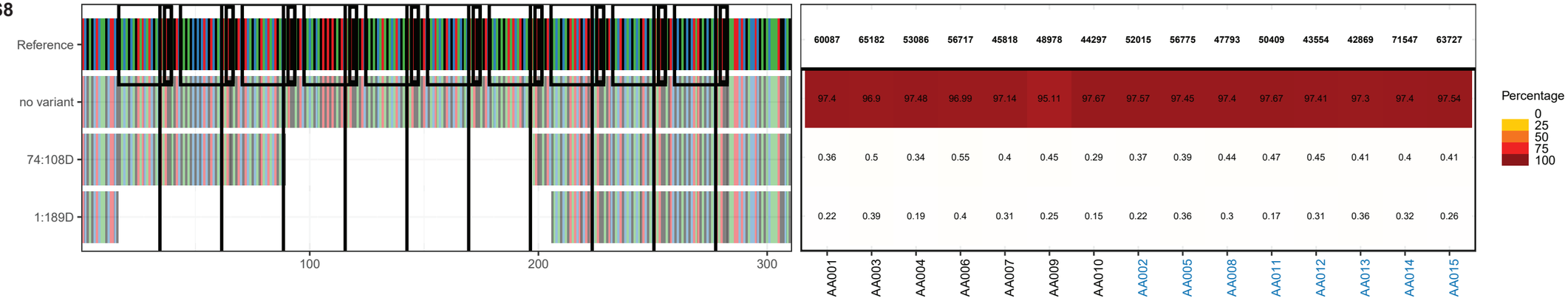
Supplementary Figure S6



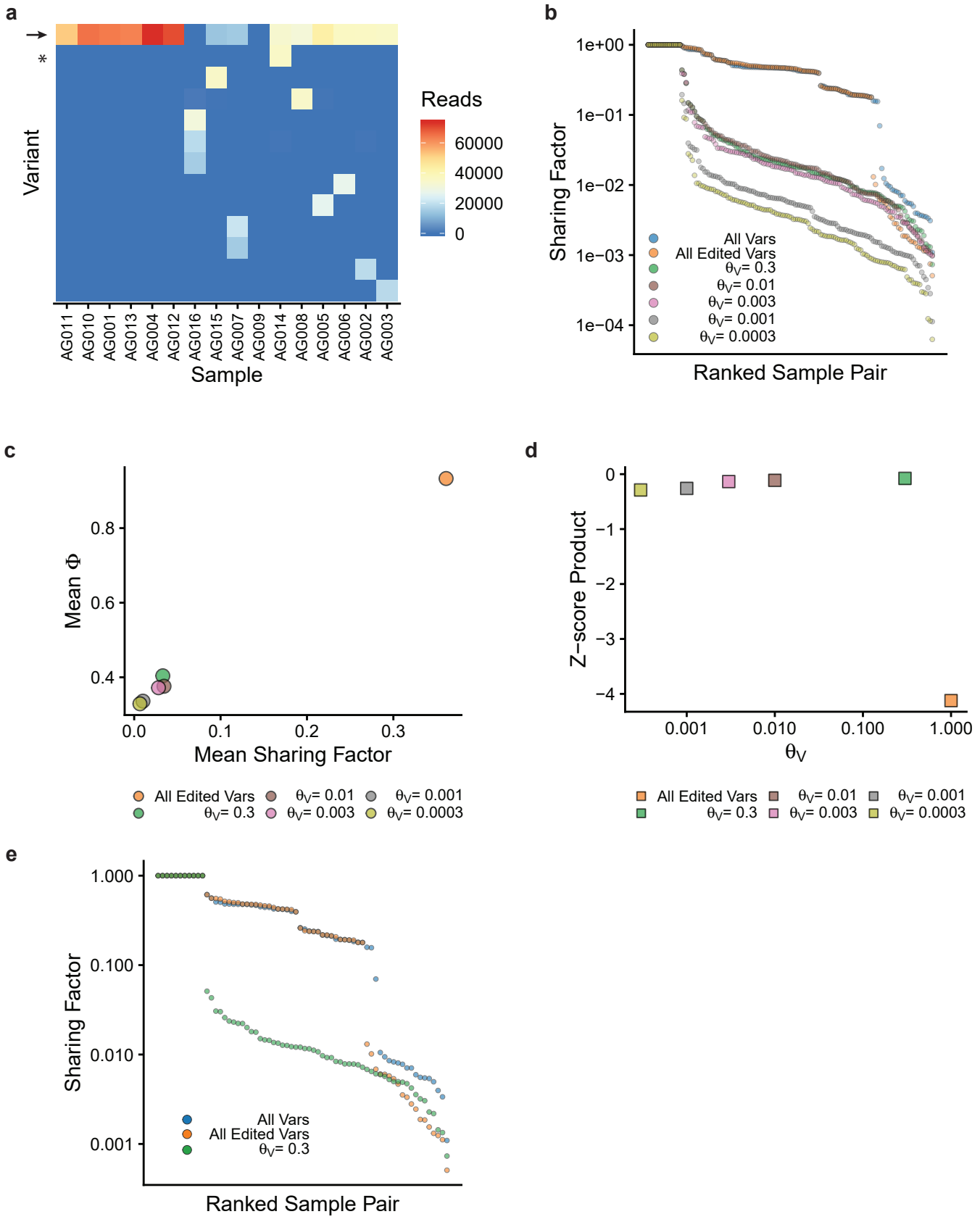
Supplementary Figure S7



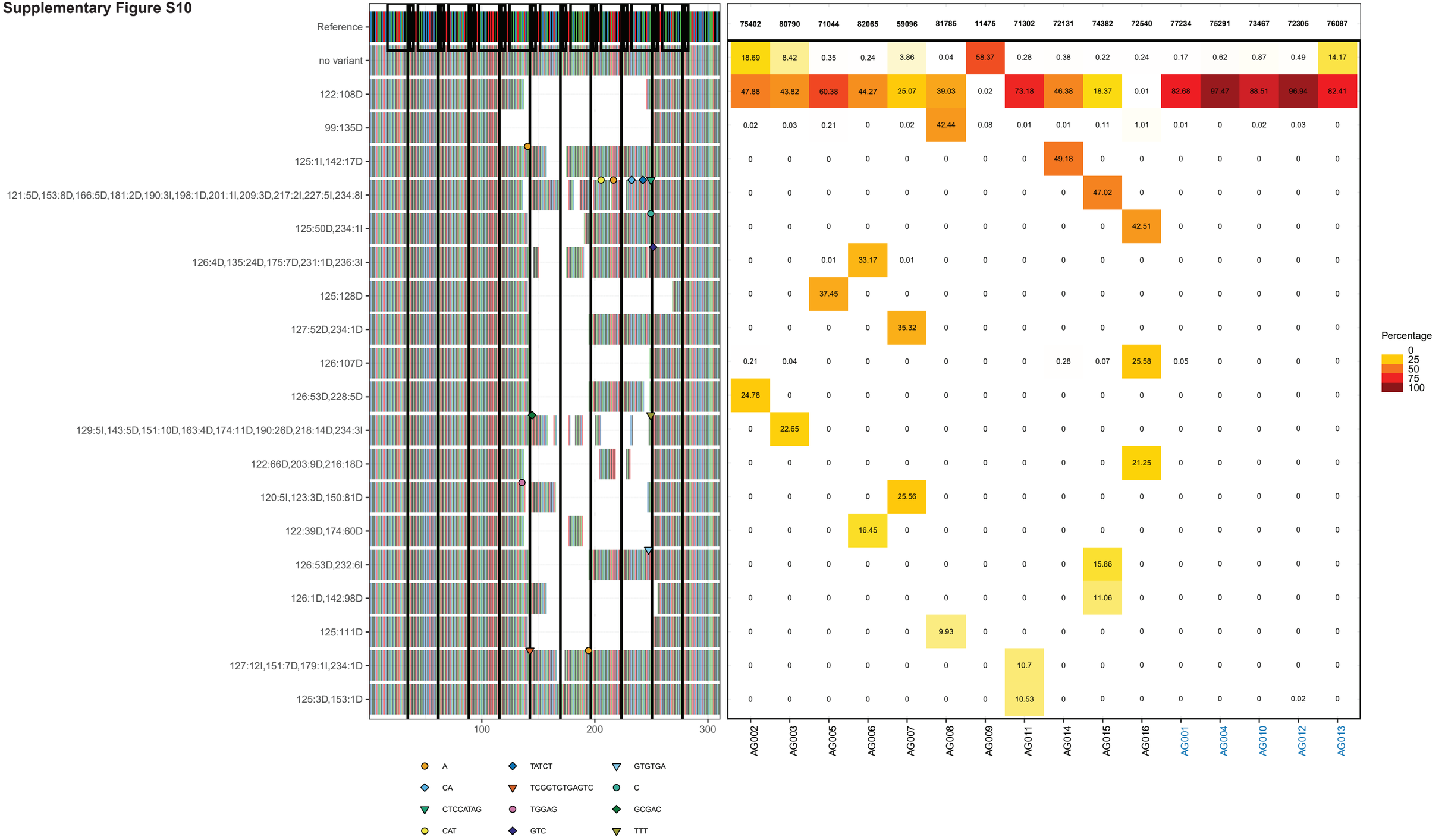
Supplementary Figure S8



Supplementary Figure S9



Supplementary Figure S10



Supplementary Figure Legends

Supplementary Figure S1. Threshold optimization and identification of informative barcodes. a) CrispRVariants produces two data tables, *C* and *P*, containing variant counts and proportions, respectively, for each sample. Variants with an allele frequency in *P* that are greater than θv are marked as common variants. b) Table *C* is split into tables $C_1 \ldots C_N$ to include single samples from the dataset. c) Common variants are identified, removed from the variant list and their counts are summed to form table $C'_1 \ldots C'_N$. This process is iterated at multiple values for θv. d) The mean fraction of informative reads and mean Sharing Factor are calculated for the dataset at each value of θv. e) The Z-score product is calculated as shown in order to select the optimal value of θv. f) Bootstrapping is used to identify samples with a high fraction of informative reads and clone numbers are calculated as described in the main text. g) SABER outputs raw variant counts and calculated HSC clone numbers which can be aggregated and plotted using any software program.

Supplementary Figure S2. Read counts for training and validation datasets analyzed with the core pipeline. Box and whisker plots show median, 1st and 3rd quartiles, and 1.5 x the interquartile range.

Supplementary Figure S3. Predicted GESTALT target editing patterns. InDelphi was used to predict individual GESTALT target repair patterns after CRISPR/Cas9 editing. Left panels show predicted nucleotide sequence; right panels show predicted frequency. Variants with predicted frequency greater than 1% are shown.

Supplementary Figure S4. Shared GESTALT variants are uncommon. a) Matrix sparsity was calculated for each sample in the training set and plotted in descending rank order. b) Scatter plot

of variant allele frequencies for GESTALT variants detected in two representative samples. 55/1821 variants had non-zero allele frequencies in both samples.

Supplementary Figure S5. GESTALT variants and allele frequencies in the training set. Left panel: The top 20 most common GESTALT variants are graphically represented and labeled by variant name. Vertical black lines represent predicted cut site for each target. Right panel: Heatmap of variant allele frequencies. Column headers indicate the number of reads for each sample. Column bases indicate sample names. Black text indicates samples with a high fraction of informative reads and blue text indicates excluded samples with a low fraction of informative reads as determined by SABER.

Supplementary Figure S6. GESTALT variants and allele frequencies in the validation set. Left panel: The top 20 most common GESTALT variants are graphically represented and labeled by variant name. Vertical black lines represent predicted cut site for each target. Right panel: Heatmap of variant allele frequencies. Column headers indicate the number of reads for each sample. Column bases indicate sample names. Black text indicates samples with a high fraction of informative reads and blue text indicates excluded samples with a low fraction of informative reads as determined by SABER.

Supplementary Figure S7. Example of an experiment with poor GESTALT barcoding. a) Heatmap of variant counts for each sample in the dataset. Asterisk indicates the unedited GESTALT allele. b) Sharing Factor curves at indicated values of θv. c) Mean Φ and mean Sharing Factor at the indicated values of θv. d) Z-score product plotted against θv. The optimal value for θv selected by SABER was 0.001. e) At $\theta v = 0.001$, the mean Sharing Factor was 0.51 in the selected samples.

Supplementary Figure S8. GESTALT variants and allele frequencies in an experiment with poor barcoding. The plot layout is similar to Supplementary figures S5 and S6.

Supplementary Figure S9. Example of an experiment with a high-frequency common GESTALT variant. a) Heatmap of variant counts for each sample in the dataset. Asterisk indicates the unedited GESTALT allele. Arrow indicates allele 122:108D. b) Sharing Factor curves at indicated values of θv. c) Mean Φ and mean Sharing Factor at the indicated values of θv. Note that after elimination of allele 122:108D with θv = 0.3, there is little further reduction in mean Φ or Sharing Factor. d) Z-score product plotted against θv. The optimal value selected by SABER was 0.3. e) At θv = 0.3, the mean Sharing Factor was 0.012 in the selected samples.

Supplementary Figure S10. GESTALT variants and allele frequencies in an experiment with a high-frequency common GESTALT variant. The plot layout is similar to Supplementary Figures S5, S6 and S8.

Supplementary Table

Supplementary Table S1. Primer sequences used in this study.

Name	Sequence (5'-3')
v6_7_F_illum	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGAGCTCAAGCTTCGG
v6_7_R_illum	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGCCATTTGTCTCGAGGTC
v6_7_UMI_F	CGCAGAGAGGCTCCGTGNNNNNNNNNNCTCAGATCTCGAGCTCAAGCTTCGG
v6_7_R	CTGCCATTTGTCTCGAGGTC
GC_tag	CGCAGAGAGGCTCCGTG